NVM Express Technical Errata

Errata ID	**010**
Change Date	**1/28/2014**
Affected Spec Ver.	**NVM Express 1.0 and NVM Express 1.1a**
Corrected Spec Ver.	

Submission info

Name	Company	Date
Ken Okin	HGST	12/18/2013

The word “device” does not have a clear meaning. The erratum clarifies sections that use the word “device” with more precise wording.

Description of the specification technical flaw:

Update the abbreviations in section 1.5 as shown below:

1.5 Conventions

Hardware shall return '0' for all bits and registers that are marked as reserved, and host software shall write all reserved bits and registers with the value of '0'.
Inside the register section, the following abbreviations are used:

RO	Read Only
RW	Read Write
R/W	Read Write. The value read may not be the last value written.
RWC	Read/Write '1' to clear
RWS	Read/Write '1' to set
Impl Spec	Implementation Specific – the controller has the freedom to choose its implementation.
HwInit	The default state is dependent on ~~device~~ NVM Express controller and system configuration. The value is initialized at reset, for example by an expansion ROM, or in the case of integrated devices, by a platform BIOS.

Update the second paragraph of section 2 as shown below:

This section details how the PCI Header, PCI Capabilities, and PCI Express Extended Capabilities should be constructed for an NVM Express ~~device~~ controller. The fields shown are duplicated from the appropriate PCI or PCI Express specifications. The PCI documents are the normative specifications for these registers and this section details additional requirements for an NVM Express ~~device~~ controller.

Update the second paragraph of section 2.5 as shown below:

Note: TLP poisoning is a mandatory capability for PCI Express implementations. There are optional features of TLP poisoning, such as TLP poisoning for a transmitter. When an NVM Express ~~device~~ controller has an error on a transmission to the host (e.g., error for a Read command), the error should be indicated as part of the NVM Express command status and not via TLP poisoning.

Update bit 28 in the PCI Express Device Capabilities register in section 2.5.3 as shown below:

28	RO	1b	**Function Level Reset Capability (FLRC):** A value of '1' indicates the Function supports the optional Function Level Reset mechanism. NVM Express ~~devices~~ controllers shall support Function Level Reset.

Update bits 07:04 in Figure 11 (Command Format – Admin Command Set) as shown below:

07:04	**Namespace Identifier (NSID):** This field specifies the namespace ID that this command applies to. If the namespace ID is not used for the command, then this field shall be cleared to 0h. If a command shall be applied to all namespaces ~~on the device~~ accessible by this controller, then this field shall be set to FFFFFFFFh. Unless otherwise noted, specifying an inactive namespace ID in a command that uses the namespace ID shall cause the controller to abort the command with status Invalid Field in Command. Specifying an invalid namespace ID in a command that uses the namespace ID shall cause the controller to abort the command with status Invalid Namespace or Format.

Update bits 07:04 in Figure 12 (Command Format – NVM Command Set) as shown below:

07:04	**Namespace Identifier (NSID):** This field specifies the namespace that this command applies to. If the namespace is not used for the command, then this field shall be cleared to 0h. If a command shall be applied to all namespaces ~~on the device~~ accessible by this controller, then this value shall be set to FFFFFFFFh.

Update bits 07:04 in Figure 13 (Command Format – Admin and NVM Vendor Specific Commands) as shown below:

07:04	**Namespace Identifier (NSID):** This field indicates the namespace ID that this command applies to. If the namespace ID is not used for the command, then this field shall be cleared to 0h. If a command shall be applied to all namespaces ~~on the device~~ accessible by this controller, then this field shall be set to FFFFFFFFh. The behavior of a controller in response to an inactive namespace ID for a vendor specific command is vendor specific. Specifying an invalid namespace ID in a command that uses the namespace ID shall cause the controller to abort the command with status Invalid Namespace or Format.

Update the first paragraph of section 4.3 as shown below:

A physical region page (PRP) entry is a pointer to a physical memory page. PRPs are used as a scatter/gather mechanism for data transfers between the controller and system memory. To enable efficient out of order data transfers between the ~~device~~ controller and the host, PRP entries are a fixed size.

Update the first two paragraphs of section 4.4.1 as shown below:

Figure 24 shows an example of a data read request using SGLs. In the example, the logical block size is 512B. The total length of the logical blocks accessed ~~on the device~~ is 13KB, of which only 11KB is transferred to the host. The Number of Logical Blocks (NLB) field in the command shall specify 26, indicating the total length of the logical blocks accessed on the device is 13KB. There are three SGL segments describing the locations in host memory where the logical block data is transferred.

The three SGL segments contain a total of three Data Block descriptors with lengths of 3 KB, 4 KB and 4 KB respectively. Segment 1 of the Destination SGL contains a Bit Bucket descriptor with a length of 2 KB that specifies to not transfer (i.e., ignore) 2 KB of logical block data from the ~~the device~~ NVM. Segment 1 of the destination SGL also contains a Last Segment descriptor specifying that the segment pointed to by the descriptor is the last SGL segment.

Update status code value 06h in Figure 30 as shown below:

06h	**Internal ~~Device~~ Error:** The command was not completed successfully due to an internal ~~device~~ error. Details on the internal device error are returned as an asynchronous event. Refer to section 5.2.

Update error status values 03h and 04h in Figure 44 as shown below:

3h	**Persistent Internal ~~Device~~ Error:** A failure occurred ~~within the controller~~ that is persistent ~~or~~ and the ~~device~~ controller is unable to isolate to a specific set of commands. If this error is indicated, then the CSTS.CFS bit may be set to '1' and the host should perform a reset as described in section 7.3.
4h	**Transient Internal ~~Device~~ Error:** A transient error occurred ~~within the device~~ that is specific to a particular set of commands; ~~and~~ controller operation may continue without a reset.

Update health status value 0h in Figure 45 as shown below:

0h	**~~Device~~ NVM subsystem Reliability:** ~~Device~~ NVM subsystem reliability has been compromised. This may be due to significant media errors, an internal error, the media being placed in read only mode, or a volatile memory backup device failing.

Update the first paragraph of section 5.10.1.1 as shown below:

This log page is used to describe extended error information for a command that completed with error or report an error that is not specific to a particular command. Extended error information is provided when the More (M) bit is set to '1' in the Status Field for the completion queue entry associated with the command that completed with error or as part of an asynchronous event with an Error status type. This log page is global to the ~~device~~ controller.

Update the second and third paragraph of section 5.10.1.2 as shown below:

Critical warnings regarding the health of the ~~device~~ NVM subsystem may be indicated via an asynchronous event notification to the host. The warnings that results in an asynchronous event notification to the host are configured using the Set Features command; refer to section 5.12.1.11.

Performance ~~of the device~~ may be calculated using parameters returned as part of the SMART / Health Information log. Specifically, the number of Read or Write commands, the amount of data read or written, and the amount of controller busy time enables both I/Os per second and bandwidth to be calculated.

Update Figure 75 as shown below:

Figure 75: Get Log Page – SMART / Health Information Log

<table>
<tr><th>Bytes</th><th>Description</th></tr>
<tr><td>0</td><td><b>Critical Warning:</b> This field indicates critical warnings for the state of the controller. Each bit corresponds to a critical warning type; multiple bits may be set. If a bit is cleared to ‘0’, then that critical warning does not apply. Critical warnings may result in an asynchronous event notification to the host.
<table>
<tr><th>Bit</th><th>Definition</th></tr>
<tr><td>00</td><td>If set to ‘1’, then the available spare space has fallen below the threshold.</td></tr>
<tr><td>01</td><td>If set to ‘1’, then the temperature has exceeded a critical threshold.</td></tr>
<tr><td>02</td><td>If set to ‘1’, then the ~~device~~ NVM subsystem reliability has been degraded due to significant media related errors or any internal error that degrades ~~device~~ NVM subsystem reliability.</td></tr>
<tr><td>03</td><td>If set to ‘1’, then the media has been placed in read only mode.</td></tr>
<tr><td>04</td><td>If set to ‘1’, then the volatile memory backup device has failed. This field is only valid if the controller has a volatile memory backup solution.</td></tr>
<tr><td>07:05</td><td>Reserved</td></tr>
</table></td></tr>
<tr><td>2:1</td><td><b>Temperature:</b> Contains the temperature of the overall ~~device~~ NVM subsystem (controller and NVM included) in units of Kelvin. If the temperature exceeds the temperature threshold, refer to section <b>Error! Reference source not found.</b>, then an asynchronous event completion may occur.</td></tr>
<tr><td>3</td><td><b>Available Spare:</b> Contains a normalized percentage (0 to 100%) of the remaining spare capacity available.</td></tr>
<tr><td>4</td><td><b>Available Spare Threshold:</b> When the Available Spare falls below the threshold indicated in this field, an asynchronous event completion may occur. The value is indicated as a normalized percentage (0 to 100%).</td></tr>
<tr><td>5</td><td><b>Percentage Used:</b> Contains a vendor specific estimate of the percentage of ~~device~~ NVM subsystem life used based on the actual ~~device~~ usage and the manufacturer’s prediction of ~~device~~ NVM life. A value of 100 indicates that the estimated endurance of the ~~device~~ NVM in the NVM subsystem has been consumed, but may not indicate ~~a device~~ an NVM subsystem failure. The value is allowed to exceed 100. Percentages greater than 254 shall be represented as 255. This value shall be updated once per power-on hour (when the controller is not in a sleep state).

Refer to the JEDEC JESD218 standard for SSD device life and endurance measurement techniques.</td></tr>
<tr><td>31:6</td><td>Reserved</td></tr>
<tr><td>47:32</td><td><b>Data Units Read:</b> Contains the number of 512 byte data units the host has read from the controller; this value does not include metadata. This value is reported in thousands (i.e., a value of 1 corresponds to 1000 units of 512 bytes read) and is rounded up. When the LBA size is a value other than 512 bytes, the controller shall convert the amount of data read to 512 byte units.

For the NVM command set, logical blocks read as part of Compare and Read operations shall be included in this value.</td></tr>
<tr><td>63:48</td><td><b>Data Units Written:</b> Contains the number of 512 byte data units the host has written to the controller; this value does not include metadata. This value is reported in thousands (i.e., a value of 1 corresponds to 1000 units of 512 bytes written) and is rounded up. When the LBA size is a value other than 512 bytes, the controller shall convert the amount of data written to 512 byte units.

For the NVM command set, logical blocks written as part of Write operations shall be included in this value. Write Uncorrectable commands shall not impact this value.</td></tr>
<tr><td>79:64</td><td><b>Host Read Commands:</b> Contains the number of read commands completed by the controller.

For the NVM command set, this is the number of Compare and Read commands.</td></tr>
</table>

95:80	**Host Write Commands:** Contains the number of write commands completed by the controller. For the NVM command set, this is the number of Write commands.
111:96	**Controller Busy Time:** Contains the amount of time the controller is busy with I/O commands. The controller is busy when there is a command outstanding to an I/O Queue (specifically, a command was issued via an I/O Submission Queue Tail doorbell write and the corresponding completion queue entry has not been posted yet to the associated I/O Completion Queue). This value is reported in minutes.
127:112	**Power Cycles:** Contains the number of power cycles.
143:128	**Power On Hours:** Contains the number of power-on hours. This does not include time that the controller was powered and in a low power state condition.
159:144	**Unsafe Shutdowns:** Contains the number of unsafe shutdowns. This count is incremented when a shutdown notification (CC.SHN) is not received prior to loss of power.
175:160	**Media Errors:** Contains the number of occurrences where the controller detected an unrecovered data integrity error. Errors such as uncorrectable ECC, CRC checksum failure, or LBA tag mismatch are included in this field.
191:176	**Number of Error Information Log Entries:** Contains the number of Error Information log entries over the life of the controller.
511:192	Reserved

Update the first paragraph of section 5.10.1.3 as shown below:

This log page is used to describe the firmware revision stored in each firmware slot supported. The firmware revision is indicated as an ASCII string. The log page also indicates the active slot number. The log page returned is defined in Figure 76. This log page is global to the ~~device~~ controller.

Update byte 260 in Figure 82 (Identify Controller Data Structure) as shown below:

260	M	**Firmware Updates (FRMW):** This field indicates capabilities regarding firmware updates. Refer to section 8.1 for more information on the firmware update process. Bits 7:4 are reserved. Bits 3:1 indicate the number of firmware slots that the ~~device~~ controller supports. This field shall specify a value between one and seven, indicating that at least one firmware slot is supported and up to seven maximum. This corresponds to firmware slots 1 through 7. Bit 0 if set to ‘1’ indicates that the first firmware slot (slot 1) is read only. If cleared to ‘0’ then the first firmware slot (slot 1) is read/write. Implementations may choose to have a baseline read only firmware image.

Update bytes 23:16 in Figure 84 (Identify Namespace Data Structure, NVM Command Set Specific) as shown below:

23:16	M	**Namespace Utilization (NUSE):** This field indicates the current number of logical blocks allocated in the namespace. This field is smaller than or equal to the Namespace Capacity. The number of logical blocks is based on the formatted LBA size. When using the NVM command set: A logical block is allocated when it is written with a Write or Write Uncorrectable command. A logical block may be deallocated using the Dataset Management command. A ~~device~~ controller may report NUSE equal to NCAP at all times if the product is not targeted for thin provisioning environments.

Update section 5.12.1.4 as shown below:

This Feature indicates the threshold for the temperature of the ~~overall device (~~controller and NVM ~~included)~~ in units of Kelvin. If this temperature is exceeded, then an asynchronous event may be issued to the host. The host should configure this feature prior to enabling asynchronous event notification for the temperature exceeding the threshold. The attributes are indicated in Command Dword 11.

If a Get Features command is submitted for this Feature, the attributes specified in Figure 95 are returned in Dword 0 of the completion queue entry for that command.

Figure 95: Temperature Threshold – Command Dword 11

Bit	Description
31:16	Reserved
15:00	**Temperature Threshold (TMPTH):** Indicates the threshold for the temperature of the ~~overall device (~~controller and NVM ~~included)~~ in units of Kelvin.

Update the fourth paragraph of section 5.12.1.8 as shown below:

This Feature is valid when the ~~device~~ controller is configured for Pin Based, MSI, Multiple MSI or MSI-X interrupts.   There is no requirement for the ~~device~~ controller to persist these settings if interrupt modes are changed. It is recommended that the host re-issue this Feature after changing interrupt modes.

Update section 5.13 as shown below:

The Format NVM command is used to low level format the NVM media. This is used when the host wants to change the LBA data size and/or metadata size. A low level format may destroy all data and metadata associated with all namespaces or only the specific namespace associated with the command (refer to the Format NVM Attributes field in the Identify Controller data structure).

As part of the Format NVM command, the host may request a secure erase of the contents of the NVM. There are two types of secure erase. The User Data Erase erases all user content present in the NVM subsystem. The Cryptographic Erase erases all user content present in the NVM subsystem by deleting the encryption key with which the user data was previously encrypted. The secure erase functionality may apply to all namespaces in the NVM subsystem including those not accessible by the controller or may be specific to a particular namespace, refer to the Identify Controller data structure.

The Format NVM command shall fail if the controller is in an invalid security state. See the TCG SIIS reference. The Format NVM command may fail if there are outstanding IO commands to the namespace specified to be formatted.

The settings specified in the Format NVM command are reported as part of the Identify Namespace data structure.

If the controller supports multiple namespaces, then the host may specify the value of FFFFFFFFh for the namespace in order to apply the format operation to all namespaces accessible by the controller regardless of the value of the Format NVM Attribute field in the Identify Controller data structure.

The Format NVM command uses the Command Dword 10 field. All other command specific fields are reserved.

Update the first paragraph of section 6 as shown below:

~~The device~~ An NVM subsystem is comprised of some number of controllers, where each controller ~~is comprised of~~ may access some number of namespaces, where each namespace is comprised of some number of logical blocks. A logical block is the smallest unit of data that may be read or written from the controller. The logical block data size, reported in bytes, is always a power of two. Logical block sizes may be 512 bytes, 1KB, 2KB, 4KB, 8KB, etc. Supported logical block sizes are reported in the Identify Namespace data structure.

Update the first paragraph of section 6.6 as shown below:

The Dataset Management command is used by the host to indicate attributes for ranges of logical blocks. This includes attributes like frequency that data is read or written, access size, and other information that may be used to optimize performance and reliability. This command is advisory; a compliant ~~device~~ controller may choose to take no action based on information provided.

Update the attributes of the Compare command in section 7.2.5.2 as shown below:

The attributes of the Compare command are:

- CMD0.CDW0.OPC is set to 05h for Compare.
- CMD0.CDW0.FUSE is set to 01b indicating that this is the first command of a fused operation.
- CMD0.CDW0.CID is set to a free command identifier.
- CMD0.CDW1.NSID is set to the appropriate namespace.
- If metadata is being used in a separate buffer, then the location of that buffer is specified.
  - If a command uses PRPs then CMD0.MPTR is set to the address of the metadata buffer.
  - If a command uses SGLs then CMD0.MSGLP is set to an SGL segment that describes the metadata buffer.
- The physical address of the first page of the data to compare.
  - If PRPs are used, CMD0.PRP1 is set to the physical address of the first page of the data to compare~~.~~ and CMD0.PRP2 is set to the physical address of the PRP List. The PRP List is shown in Figure 176 for a PRP List with three entries.
  - If the command uses SGLs, CMD0.SGL1 is set to an appropriate SGL segment descriptor depending on whether more than one descriptor is needed.
- CMD0.CDW10.SLBA is set to the first LBA to compare against. Note that this field also spans Command Dword 11.
- CMD0.CDW12.LR is set to ‘0’ to indicate that the controller should apply all available error recovery means to retrieve the data for comparison.
- CMD0.CDW12.FUA is cleared to ‘0’, indicating that the data may be read from any location, including a DRAM cache, on the ~~device~~ NVM subsystem.
- CMD0.CDW12.PRINFO is cleared to 0h since end-to-end protection is not enabled.
- CMD0.CDW12.NLB is set to 3h, indicating that four logical blocks of a size of 4KB each are to be compared against.
- CMD0.CDW14 is cleared to 0h since end-to-end protection is not enabled.
- CMD0.CDW15 is cleared to 0h since end-to-end protection is not enabled.

Update the attributes of the Write command in section 7.2.5.2 as shown below:

The attributes of the Write command are:

- CMD1.CDW0.OPC is set to 01h for Write.
- CMD1.CDW0.FUSE is set to 10b indicating that this is the second command of a fused operation.
- CMD1.CDW0.CID is set to a free command identifier.
- CMD1.CDW1.NSID is set to the appropriate namespace. This value shall be the same as CMD0.CDW1.NSID.
- If metadata is being used in a separate buffer, then the location of that buffer is specified.

    - If a command uses PRPs then CMD1.MPTR is set to the address of the metadata buffer.
    - If a command uses SGLs then CMD1.MSGLP is set to an SGL segment that describes the metadata buffer.
- The physical address of the first page of data to write is identified.
    - If the command uses PRPs, then CMD1.PRP1 is set to the physical address of the first page of the data to write~~.~~ and CMD1.PRP2 is set to the physical address of the PRP List. The PRP List includes three entries.
    - If the command uses SGLs, CMD0.SGL1 is set to an appropriate SGL segment descriptor depending on whether more than one descriptor is needed.
- CMD1.CDW10.SLBA is set to the first LBA to compare against. Note that this field also spans Command Dword 11. This value shall be the same as CMD0.CDW10.SLBA.
- CMD1.CDW12.LR is set to '0' to indicate that the controller should apply all available error recovery means to write the data to the NVM.
- CMD1.CDW12.FUA is cleared to '0', indicating that the data may be written to any location, including a DRAM cache, on the ~~device~~ NVM subsystem.
- CMD1.CDW12.PRINFO is cleared to 0h since end-to-end protection is not enabled.
- CMD1.CDW12.NLB is set to 3h, indicating that four logical blocks of a size of 4KB each are to be compared against. This value shall be the same as CMD0.CDW12.NLB.
- CMD1.CDW14 is cleared to 0h since end-to-end protection is not enabled.
- CMD1.CDW15 is cleared to 0h since end-to-end protection is not enabled.

Update the first paragraph of section 8.4.1 as shown below:

The controller may support autonomous power state transitions, as indicated in the Identify Controller data structure in Figure 82. Autonomous power state transitions provide a mechanism for the host to configure the ~~device~~ controller to automatically transition between power states on certain conditions without software intervention.

Update section 9.4 as shown below:

9.4 Internal Controller ~~Device~~ Error Handling

~~Device specific errors~~ Errors such as a DRAM failure or power loss notification ~~errors~~ indicate that a controller level failure has occurred during the processing of a command. The status code of the completion queue entry should indicate an Internal ~~Device~~ Error status code (if multiple error conditions exist, the lowest numerical value is returned). Host software shall ignore any data transfer associated with the command. The host may choose to re-submit the command or indicate an error to the higher level software.

Disposition log

12/18/2013	Erratum captured.
12/20/2013	Edits based on 12/19 meeting.
1/22/2014	Proposal for fields in Figure 75 to be per controller or per subsystem.
1/28/2014	Removed proposal for Figure 75 on controller/subsystem, will be addressed in TP.
3/17/2014	Erratum ratified.